

# Jointly interventional and observational data: estimation of interventional Markov equivalence classes of directed acyclic graphs

Alain Hauser and Peter Bühlmann  
`{hauser, buhlmann}@stat.math.ethz.ch`  
 Seminar für Statistik  
 ETH Zürich  
 8092 Zürich, Switzerland

## Abstract

In many applications we have both observational and (randomized) interventional data. We propose a Gaussian likelihood framework for joint modeling of such different data-types, based on global parameters consisting of a directed acyclic graph (DAG) and corresponding edge weights and error variances. Thanks to the global nature of the parameters, maximum likelihood estimation is reasonable with only one or few data points per intervention. We prove consistency of the BIC criterion for estimating the interventional Markov equivalence class of DAGs which is smaller than the observational analogue due to increased partial identifiability from interventional data. Such an improvement in identifiability has immediate implications for tighter bounds for inferring causal effects. Besides methodology and theoretical derivations, we present empirical results from real and simulated data.

Keywords: Causal inference; Interventions; BIC; Graphical model; Maximum likelihood estimation; Greedy equivalence search

## 1 Introduction

Causal inference often relies on an underlying influence diagram in terms of a directed acyclic graph (DAG). In absence of knowledge of the true underlying DAG, there has been a substantial line of research to estimate the Markov equivalence class of DAGs which is identifiable from data. Most often, the target of interest is the observational Markov equivalence class to be inferred from observational data; that is, the data arises from observing a system in “steady state” without any interventions, see for example the books by Spirtes et al. (2000) or Pearl (2000). For the important case of multivariate Gaussian distributions, the observational Markov equivalence class is rather large and thus, many parts of the true underlying DAG are unidentifiable from observational data, see for example Verma and Pearl (1990) or Andersson et al. (1997) for a graphical characterization of the Markov equivalence class in the Gaussian or the fully nonparametric case. Under additional assumptions, identifiability of the whole DAG is guaranteed as with linear structural equation models with non-Gaussian errors (Shimizu et al., 2006) or additive noise models (Hoyer et al., 2009), see also Peters et al. (2011).

In many applications, we have both observational and interventional data, where the latter are coming from (randomized) intervention experiments. In biology, for example, we often have observational data from a wildtype individual and interventional data from mutants or individuals with knocked-out genes. Besides the methodological issue of properly modeling such data, we gain in terms of identifiability: the interventional Markov equivalence class is smaller (Hauser and Bühlmann, 2012), thanks to additional interventional experiments, and this is of

particular interest for the Gaussian and nonparametric cases which are hardest in terms of identifiability.

We focus here on the problem of joint modeling of observational and interventional Gaussian data. Thereby, we assume that the observational distribution is Markovian (and typically faithful; cf. Spirtes et al., 2000) to a true underlying DAG  $D_0$  and that the different interventional distributions are linked to the DAG  $D_0$  and the observational distribution via the intervention calculus using the do-operator (Pearl, 2000). Linking all interventional distributions to the same DAG  $D_0$  and the single observational distribution allows to deal with the situation where we have only one interventional data point for every intervention target (intervention experiment). We propose to use the maximum likelihood estimator which has not been studied or even used for the observational-interventional data setting. We prove that when penalizing with the BIC score, it consistently identifies the true underlying observational-interventional Markov equivalence class.

### 1.0.1 Relation to other work

Some approaches to incorporate interventional data for learning causal models have been developed in earlier work. Cooper and Yoo (1999) and Eaton and Murphy (2007) address the problem of calculating a posterior (and also a likelihood) of a data set having observational as well as interventional data but do not investigate properties of the Bayesian estimators e.g. in the large-sample limit nor address the issue of identifiability or Markov equivalence. He and Geng (2008) present a method which first estimates the observational Markov equivalence class and then in a second step, it identifies additional structure using interventional data. This technique is inefficient due to decoupling into two stages, especially if one has many interventional but only a few observational data: in fact, our maximum likelihood estimator in Section 3 can cope with the situation where we have interventional data only. To our knowledge, no analysis of the maximum likelihood estimator of an ensemble of observational and interventional data has been pursued so far. The computation of the maximum likelihood estimator which we will briefly indicate in Section 4.2 has been developed in Hauser and Bühlmann (2012): due to its non-trivial nature, it is not dealt with in this paper. When having observational data only, the works by Chickering (2002a,b) are dealing with maximum likelihood estimation and consistency of the BIC score for the corresponding observational Markov equivalence class: however, the extension to the mixed interventional-observational case, which occurs in many real problems, is a highly non-trivial step.

## 2 Interventional-observational data and maximum likelihood estimation

We start by presenting the model and the corresponding maximum likelihood estimator.

### 2.1 A Gaussian model

We consider the setting with  $n_{\text{obs}}$  observational and  $n_{\text{int}}$  interventional  $p$ -variate data from the following model:

$$\begin{aligned} X^{(1)}, \dots, X^{(n_{\text{obs}})} &\text{ i.i.d. } \sim P_{\text{obs}}, \\ X_{\text{int}}^{(1)}, \dots, X_{\text{int}}^{(n_{\text{int}})} &\text{ independent, and independent of } X^{(1)}, \dots, X^{(n_{\text{obs}})}, \quad X_{\text{int}}^{(i)} \sim P_{\text{int}}^{(i)}. \end{aligned} \quad (1)$$

In the following, we specify the observational distribution  $P_{\text{obs}}$  and all the interventional distributions  $P_{\text{int}}^{(i)}$  ( $i = 1, \dots, n_{\text{int}}$ ).

Regarding the observational distribution, we assume that

$$P_{\text{obs}} = \mathcal{N}_p(0, \Sigma), \text{ where } P_{\text{obs}} \text{ is Markovian with respect to a DAG } D. \quad (2)$$

The assumption with mean zero is not really a restriction: all derivations can be easily adapted, at the price of writing an intercept in many formulas. An implementation in the R-package `pcalg` (Kalisch et al., 2012) offers the option to restrict to mean zero or not. The Markovian assumption is equivalent to the factorization property in (3) below. We sometimes refer to the true observational distribution as  $P_{0,\text{obs}}$  with parameter  $\Sigma_0$ , and the true DAG is  $D_0$ .

In the following, the set of nodes in a DAG  $D$ , associated to the  $p$ -dimensional random vector  $(X_1, \dots, X_p)$ , is denoted by  $\{1, \dots, p\}$  and the parental set by

$$\text{pa}(j) = \text{pa}_D(j) = \{k; k \text{ a parent of node } j\} \quad (j = 1, \dots, p) .$$

The Markov condition of  $P_{\text{obs}}$  with respect to the DAG  $D$ , with parental sets  $\text{pa}(\cdot) = \text{pa}_D(\cdot)$ , allows the following (minimal) factorization of the joint distribution (Lauritzen, 1996):

$$f_{\text{obs}}(x) = \prod_{j=1}^p f_{\text{obs}}(x_j | x_{\text{pa}(j)}), \quad (3)$$

where  $f_{\text{obs}}(\cdot)$  denotes the Gaussian density of  $P_{\text{obs}}$  and  $f_{\text{obs}}(x_j | x_{\text{pa}(j)})$  are univariate Gaussian conditional densities.

The interventional distributions  $P_{\text{int}}^{(i)}$  ( $i = 1, \dots, n_{\text{int}}$ ) may all be different but linked to the same observational distribution  $P_{\text{obs}}$  and the same DAG  $D$  via the intervention calculus in Section 2.1.1. Due to the common underlying model given by  $P_{\text{obs}}$  and the DAG  $D$ , this allows to handle cases where we have only one interventional data point for every interventional distribution.

### 2.1.1 Intervention calculus

The intervention calculus, or do-calculus (Pearl, 2000), is a key concept for describing the model of the intervention distributions. We consider the DAG  $D$  appearing in the observational model (2), and we assign it a *causal* interpretation as follows. Assume  $X_{\text{int}}$  is realized under a (single- or multi-variable) intervention at the intervention target  $I \subseteq \{1, \dots, p\}$  denoting the set of intervened vertices. The distribution of  $X_{\text{int}}$  is then given by the so-called truncated factorization, a version of the factorization in (3). The truncated factorization for the interventional distribution for  $X_{\text{int}}$  with deterministic intervention  $\text{do}(X_I = u_I)$  is defined as (Pearl, 2000):

$$f_{\text{int}}(x_{I^c} | \text{do}(X_I = u_I)) = \prod_{j \notin I} f_{\text{obs}}(x_j | x_{\text{pa}(j) \cap I^c}, u_{\text{pa}(j) \cap I}),$$

where  $f_{\text{int}}(\cdot | \text{do}(X_I = u_I))$  is the intervention Gaussian density when doing an intervention at  $X_I$  by setting it to the value  $u_I$ , and  $f_{\text{obs}}(\cdot | \cdot)$  is as in (3). Here, the conditioning argument  $x_{\text{pa}(j) \cap I^c}, u_{\text{pa}(j) \cap I}$  distinguishes the value of the unintervened variables  $x_{\text{pa}(j) \cap I^c}$  and the values of the intervened variables  $u_{\text{pa}(j) \cap I}$ .

Deterministic interventions as described above make the intervened variables  $X_I$  deterministic, having the value of the intervention levels  $u_I$ . In this paper, we consider stochastic interventions where the intervened variables  $X_I$  are set to the value of a random vector  $U_I \sim \prod_{j \in I} f_{U_j}(u_j) du_j$  with independent (but not necessarily identically distributed) components having densities  $f_{U_j}(\cdot)$  ( $j \in I$ ). The truncated factorization for stochastic interventions (where the intervention values are independent of the observational variables) then reads as follows:

$$f_{\text{int}}(x | \text{do}(X_I = U_I)) = \prod_{j \notin I} f_{\text{obs}}(x_j | x_{\text{pa}(j) \cap I^c}, U_{\text{pa}(j) \cap I}) \prod_{j \in I} f_{U_j}(x_j). \quad (4)$$

In contrast to the case of deterministic interventions above, the intervention density (4) is  $p$ -variate:  $x \in \mathbb{R}^p$  and for  $j \in I$ ,  $x_j$  is then an argument in the density from the random

intervention variable  $U_j$ . In the following, we assume that the densities for the intervention values are Gaussian as well:

$$U_1, \dots, U_p \text{ independent with } U_j \sim \mathcal{N}(\mu_{U_j}, \tau_j^2) \ (j = 1, \dots, p). \quad (5)$$

The truncated factorization in (4) or its deterministic version above can be obtained by applying the Markov property to the interventional DAG  $D_I$ : given a DAG  $D$ , the intervention DAG  $D_I$  is defined as  $D$  but deleting all directed edges which point into  $i \in I$ , for all  $i \in I$ .

An interventional data point  $X_{\text{int}}^{(i)}$ , with intervention target  $T^{(i)} = I \subseteq \{1, \dots, p\}$  and corresponding intervention value  $U_I^{(i)}$ , then has density  $f_{\text{int}}(x | \text{do}(X_I^{(i)} = U_I^{(i)}))$  from (4). Thus, in other words, the intervention distribution  $P_{\text{int}}^{(i)}$  is characterized by the Gaussian density in (4). This, together with the specific form of the Gaussian observational distribution (see also (3)), fully specifies the model in (1) which then reads as:

$$\begin{aligned} &X_{\text{obs}}^{(1)}, \dots, X_{\text{obs}}^{(n_{\text{obs}})} \text{ i.i.d. } \sim f_{\text{obs}}(x)dx \text{ as in (3),} \\ &X_{\text{int}}^{(1)}, \dots, X_{\text{int}}^{(n_{\text{int}})} \text{ independent, and independent of } X_{\text{obs}}^{(1)}, \dots, X_{\text{obs}}^{(n_{\text{obs}})}, \\ &X_{\text{int}}^{(i)} \sim f_{\text{int}}(x | \text{do}(X_{T^{(i)}} = U_{T^{(i)}}^{(i)}))dx \text{ as in (4)} \\ &U^{(1)}, \dots, U^{(n_{\text{int}})} \text{ independent, and independent of } X_{\text{obs}}^{(1)}, \dots, X_{\text{obs}}^{(n_{\text{obs}})}, \\ &U^{(i)} \sim \mathcal{N}\left(\mu_U^{(i)}, \text{diag}(\tau_1^{(i)2}, \dots, \tau_p^{(i)2})\right) \end{aligned} \quad (6)$$

The true underlying parameters and quantities in the model (6) are denoted by  $\mu_0, \Sigma_0, \mu_{0,U}^{(i)}, \{\tau_{0,j}^{(i)2}\}_j$  and the true DAG  $D_0$ . It is well known, see also Section 3, that  $D_0$  is typically not identifiable from the observational and a few interventional distributions.

### 2.1.2 Structural equation model

The model in (6) (or in (1)) can be alternatively written as a linear structural equation model thanks to the Gaussian assumption. The observational variables can be represented as

$$X_{\text{obs},k} = \sum_{j=1}^p \beta_{kj} X_{\text{obs},j} + \varepsilon_k, \ \varepsilon_k \sim \mathcal{N}(0, \sigma_k^2) \ (k = 1, \dots, p), \quad (7)$$

where  $\beta_{kj} = 0$  if  $j \notin \text{pa}(k) = \text{pa}_D(k)$  and  $\varepsilon_1, \dots, \varepsilon_n$  are independent and  $\varepsilon_k$  independent of  $X_{\text{obs}, \text{pa}(k)}$ . Using the matrix  $B = (\beta_{kj})_{k,j=1}^p$  with

$$B \in \mathbf{B}(D) := \{A = (\alpha_{kj}) \in \mathbb{R}^{p \times p}; \ \alpha_{kj} = 0 \text{ if } j \notin \text{pa}_D(k)\}, \quad (8)$$

we can write

$$X_{\text{obs}} = BX_{\text{obs}} + \varepsilon, \ \varepsilon \sim \mathcal{N}_p(0, \text{diag}(\sigma_1^2, \dots, \sigma_p^2)).$$

An interventional setting with intervention  $\text{do}(X_I = U_I)$  (and intervention target  $T = I$ ) can be represented as follows:

$$X_{\text{int},k} = \begin{cases} \sum_{j \notin I} \beta_{kj} X_{\text{int},j} + \sum_{j \in I} \beta_{kj} U_j + \varepsilon_k & , \text{ if } k \notin I, \\ U_k & , \text{ if } k \in I, \end{cases} \quad (9)$$

with  $\beta_{kj}$  and  $\varepsilon_k$  as in (7) with the additional property that  $U$  is independent of  $X_{\text{obs}}$  and  $\varepsilon$ .

Thus, the model in (6) is given as

$$\begin{aligned}
& X_{\text{obs}}^{(1)}, \dots, X_{\text{obs}}^{(n_{\text{obs}})} \text{ i.i.d. as in (7),} \\
& X_{\text{int}}^{(1)}, \dots, X_{\text{int}}^{(n_{\text{int}})} \text{ independent, and independent of } X^{(1)}, \dots, X^{(n_{\text{obs}})}, \\
& X_{\text{int}}^{(i)} \text{ as in (9) with intervention target } I = T^{(i)}, \\
& U^{(1)}, \dots, U^{(n_{\text{int}})} \text{ independent, and independent of } X_{\text{obs}}^{(1)}, \dots, X_{\text{obs}}^{(n_{\text{obs}})}, \\
& U^{(i)} \sim \mathcal{N}\left(\mu_U^{(i)}, \text{diag}(\tau_1^{(i)2}, \dots, \tau_p^{(i)2})\right)
\end{aligned} \tag{10}$$

It also holds that for  $\varepsilon$  in (7),  $\varepsilon^{(1)}, \dots, \varepsilon^{(n)}$  are independent of  $U^{(1)}, \dots, U^{(n)}$ . As before, we denote the true underlying quantities by  $B_0, \{\sigma_{0,k}^2\}_k, \mu_{0,U}, \tau_0^2$  and the true DAG  $D_0$ . Because of the causal interpretation of the DAG model, we call a model as in (10) or (6) a Gaussian causal model in the following.

## 2.2 Maximum likelihood estimation when the DAG is given

The likelihood for the Gaussian model (6) is parameterized by the covariance matrix  $\Sigma$  of  $P_{\text{obs}} = \mathcal{N}_p(0, \Sigma)$ , the DAG  $D$  and the parameters  $\mu_U^{(i)}, \tau^{(i)2}$  for the stochastic intervention values  $U^{(i)}$ . Alternatively, and the route taken here, we can use the linear structural equation model and parameterize the likelihood with the coefficient matrix  $B$ , the error variances  $\sigma_1^2, \dots, \sigma_p^2$ , and  $\mu_U^{(i)}, \tau^{(i)2}$ . Using this, the matrix  $B$  is constrained such that its non-zero elements are corresponding to the directed edges in the DAG  $D$ .

For a given DAG  $D$ , it is rather straightforward to derive the maximum likelihood estimator, as discussed below. Much more involved is the issue of structure learning when the DAG  $D$  is unknown: there we want to estimate a suitable Markov equivalence of the unknown DAG, as discussed in Section 3.

It is easy to see that the log-likelihood for  $\mu_{U_j}^{(i)}, \tau_j^{(i)2}$  decouples from the remaining parameters, and we regard  $\mu_{U_j}^{(i)}, \tau_j^{(i)2}$  (for all  $i$  and  $j$ ) as nuisance parameters.

In the sequel, we unify the notation and denote an observational data point with the intervention target  $I = \emptyset$ . We can then write the distribution of  $X_{\text{int}} | \text{do}(X_I = U_I)$  as:

$$\begin{aligned}
X | \text{do}(X_I = U_I) & \sim \mathcal{N}(\mu^{(I)}, \Sigma^{(I)}), \\
\mu^{(I)} & = (\mathbb{I} - R^{(I)}B)^{-1} Q^{(I)\text{T}} \mu_{U_I}, \\
\Sigma^{(I)} & = (\mathbb{I} - R^{(I)}B)^{-1} [R^{(I)} \text{diag}(\sigma^2) R^{(I)} + Q^{(I)\text{T}} \text{diag}(\tau_I^2) Q^{(I)}] (\mathbb{I} - R^{(I)}B)^{-\text{T}}.
\end{aligned} \tag{11}$$

Thereby, we have used the following matrices:

$$\begin{aligned}
P^{(I)} & : \mathbb{R}^p \rightarrow \mathbb{R}^{p-|I|}, \quad x \mapsto x_{I^c}, \\
Q^{(I)} & : \mathbb{R}^p \rightarrow \mathbb{R}^{|I|}, \quad x \mapsto x_I, \\
R^{(I)} & : \mathbb{R}^p \rightarrow \mathbb{R}^p, \quad R^{(I)} := P^{(I)\text{T}} P^{(I)}.
\end{aligned} \tag{12}$$

The Gaussian distribution in (11) is a direct consequence of (9), which can be rewritten in vector-matrix notation as

$$X_{\text{int}} = R^{(I)} (B X_{\text{int}} + \varepsilon) + Q^{(I)\text{T}} U.$$

Denoting the intervention target for the  $i$ th data point  $X^{(i)}$  by  $T^{(i)}$ , and the total sample size as  $n = n_{\text{obs}} + n_{\text{int}}$ , the log-likelihood (conditional on  $U^{(1)}, \dots, U^{(n)}$ ) becomes

$$\ell_D(B, \{\sigma_k^2\}_k, \{\mu_U^{(i)}\}_i, \{\tau^{(i)2}\}_i; T^{(1)}, \dots, T^{(n)}, X^{(1)}, \dots, X^{(n)}) = \sum_{i=1}^n \log f_{\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})}}(X^{(i)}),$$

where  $f_{\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})}}$  denotes the density of  $\mathcal{N}(\mu^{(T^{(i)})}, \Sigma^{(T^{(i)})})$  in (11) which depends on  $B$ ,  $\{\sigma_k^2\}_k$ ,  $\{\mu_U^{(i)}\}_i$  and  $\{\tau^{(i)2}\}_i$ . To make notation shorter, we will denote by  $\mathcal{T}$  the sequence of intervention targets  $T^{(1)}, \dots, T^{(n)}$  in the following, and by  $\mathbf{X}$  the data matrix consisting of the rows  $X^{(1)}$  to  $X^{(n)}$ .

For a given DAG structure  $D$ , implying certain zeroes in  $B \in \mathbf{B}(D)$  through the space  $\mathbf{B}(D)$  in (8), the maximum likelihood estimator is defined as:

$$\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k = \arg \min_{\substack{B \in \mathbf{B}(D) \\ \{\sigma_i^2\} \in (\mathbb{R}^+)^p}} -\ell_D(B, \{\sigma_i^2\}_i; \mathcal{T}, \mathbf{X}). \quad (13)$$

The expressions  $\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k$  have an explicit form as described in Section 6.1; the nuisance parameters  $\{\mu_U^{(i)}\}_i, \{\tau^{(i)2}\}_i$  do not appear in (13) any more since the minimizer of the likelihood does not depend on them.

### 3 Estimation of the interventional Markov equivalence class

Consider the model in (6) or (10). It is well known that one cannot identify the underlying DAG  $D_0$  from  $P_{\text{obs}} = P_{0,\text{obs}}$ . However, assuming e.g. faithfulness of the distribution as in (2), one can identify the observational Markov equivalence class  $\mathcal{E}(D_0) = \mathcal{E}(P_{\text{obs}})$  from  $P_{\text{obs}}$ , see for example Spirtes et al. (2000) or Pearl (2000).

#### 3.1 Characterizing the interventional Markov equivalence class

The power of interventional data is that we can identify more than the observational Markov equivalence class, namely the smaller interventional Markov equivalence classes (Hauser and Bühlmann, 2012). Regarding the latter, we consider a family of intervention targets, a subset of the powerset of the vertices  $\{1, \dots, p\}$ :  $\mathcal{I} \subset \mathcal{P}(\{1, \dots, p\})$ . In our context  $\mathcal{I} = \{T^{(i)} \subseteq \{1, \dots, p\}; i = 1, \dots, n\}$  is the set of intervention targets of the  $n_{\text{int}}$  interventional data together with the empty set  $\emptyset$  as long as we have at least one observational data point ( $n_{\text{obs}} > 0$ ).

A family of targets  $\mathcal{I}$  is called conservative if for all  $j \in \{1, \dots, p\}$ , there is some  $I \in \mathcal{I}$  such that  $j \notin I$ . The simplest such family is  $\mathcal{I} = \{\emptyset\}$ , i.e., observational data only. Furthermore, every  $\mathcal{I}$  arising from an ensemble of observational and interventional data is a conservative family of targets as well. The issue that a family of targets should be conservative is crucial for characterization of interventional Markov equivalence classes (Hauser and Bühlmann, 2012), and with having jointly observational and interventional data in mind, it is not really a restriction.

The example in Figure 1 shows three DAGs that are observationally Markov equivalent since they have the same skeleton (i.e., they yield the same undirected graph if all directed edges are replaced by undirected ones) and the same v-structures (i.e., induced subgraphs of the form  $a \rightarrow b \leftarrow c$ ) (Verma and Pearl, 1990). If we have, in addition to observational data, data from an intervention at vertex 4, the orientations of the arrows incident to the intervened vertex become identifiable. Technically speaking, the interventional Markov equivalence class under the family of targets  $\mathcal{I} = \{\emptyset, \{4\}\}$  is *smaller* than the observational Markov equivalence class.

The general definition of an interventional Markov equivalence class is given in Section 6.2. The interventional Markov equivalence class  $\mathcal{E}_{\mathcal{I}}(D_0)$  is identifiable from  $P_{0,\text{obs}}$  in (2) and the interventional distributions, given by  $f_{\text{int}}(x | \text{do}(X_I = U))dx$  in (6) for all  $I \in \mathcal{I}$ , assuming faithfulness as in assumptions (A1) and (A2) below. In Hauser and Bühlmann (2012), the interventional Markov equivalence class of a DAG  $D$  for a conservative family of intervention targets  $\mathcal{I}$  is rigorously characterized in terms of a chain graph with directed and undirected edges, the so-called interventional essential graph or  $\mathcal{I}$ -essential graph.

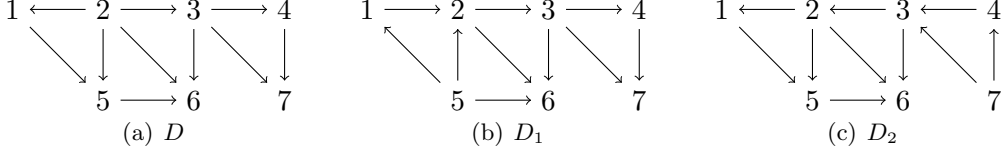


Figure 1: Three DAGs having equal skeletons and a single v-structure,  $3 \rightarrow 6 \leftarrow 5$ , hence being observationally Markov equivalent. Under the family of intervention targets  $\mathcal{I} = \{\emptyset, \{4\}\}$ ,  $D$  and  $D_1$  are still (interventionally) Markov equivalent (i.e., statistically indistinguishable), while  $D$  and  $D_2$  belong to different interventional Markov equivalence classes.

### 3.2 Structure learning using BIC-score

For estimating the structure and the parameters of the interventional Markov-equivalence class, we consider the penalized maximum-likelihood estimator using the BIC-score. Denote by  $\hat{B}(D)$  and  $\{\hat{\sigma}_k^2(D)\}_k$  the maximum-likelihood estimators for a given DAG  $D$ , as in (13). An estimate for the interventional Markov-equivalence class is then:

$$\hat{\mathcal{E}}_{\mathcal{I}} = \arg \min_{\mathcal{E}_{\mathcal{I}}(D)} -\ell_D(\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{T}, \mathbf{X}) + \frac{1}{2} \log(n) \dim(\mathcal{E}_{\mathcal{I}}(D)), \quad (14)$$

$$\dim(\mathcal{E}_{\mathcal{I}}(D)) = \dim(D) = \text{number of non-zero elements in } \hat{B}(D).$$

The optimization is over all interventional Markov equivalence classes with corresponding DAGs  $D$ , see also Section 3.3 below.

We note that the  $\ell_0$ -penalty has the property that the score remains invariant for all members in the interventional Markov equivalence class  $\mathcal{E}_{\mathcal{I}}(D)$ : this property is not true for some other penalties such as the  $\ell_1$ -norm. We outline in Section 3.3 a computational algorithm for computing the estimator in (14).

We now justify the estimator in (14) by providing a consistency result. We make the following assumptions.

- (A1) The true observational distribution  $P_{0,\text{obs}}$  in (2), or equivalently the distribution of  $X_{\text{obs}} \sim f_{\text{obs}}(x)dx$  in (6) is faithful with respect to the true underlying DAG  $D_0$ .
- (A2) The true interventional distributions of  $X_{\text{int}}^{(i)} \sim f_{\text{int}}(x | \text{do}(X_{T^{(i)}} = U^{(i)}))dx$  in (6) are faithful with respect to the true underlying intervention DAG  $D_{0,T^{(i)}}$ , for all  $i = 1, \dots, n_{\text{int}}$  (for the definition of the intervention DAG, see Section 2.1.1).

The faithfulness assumption means that all marginal and conditional independencies can be read off from the DAG, here  $D_0$  or  $D_{0,T^{(i)}}$ , respectively (Spirtes et al., 2000). This is a stronger requirement than a Markov assumption which allows to infer some conditional independencies from the DAG  $D_0$  or  $D_{0,T^{(i)}}$ .

In our case with a data set arising from different interventions, we do not have identically distributed data, as it is evident for example from equation (11). To be able to make a precise consistency statement for the estimator (14), we regard the sequence of intervention targets as *random*:

- (A3) The intervention targets  $T^{(1)}, \dots, T^{(n)}$  are  $n$  i.i.d. realizations of a random variable  $T$  taking values in  $\mathcal{I}$ :  $P[T = I] = w_I > 0$  for all  $I \in \mathcal{I}$ .

In Section 2.2, we have already seen that the parameters  $\mu_{U_j}^{(i)}, \tau_j^{(i)2}$  (for all  $i$  and  $j$ ) are nuisance parameters. They do not belong to the statistical model, but describe the experimental setting (i.e., the interventions). With assumption (A3), we introduce an additional, “artificial” set of nuisance parameters describing the experimental setting. By this approach, we can model the

sequence  $(T^{(i)}, X^{(i)})_{i=1}^n$  as independent realizations of random variables  $(T, X) \in \mathcal{I} \times \mathbb{R}^p$  with the following distribution:

$$P[T = I] = w_I, \quad f(x \mid T = I) = f_{\text{int}}(x \mid \text{do}(X_I = U_I)) .$$

**Theorem 1** *Consider model (6) with the family of intervention targets  $\mathcal{I}$ . Assume (A1), (A2) and (A3). Then: as  $n \rightarrow \infty$ ,*

$$\mathbb{P}[\hat{\mathcal{E}}_{\mathcal{I}} = \mathcal{E}_{\mathcal{I}}(D_0)] \rightarrow 1.$$

A proof is given in Section 6.3. The result might not be surprising in view of model selection consistency results of BIC for curved exponential family models (Haughton, 1988). However, a careful analysis is needed to cope with the special situation of data arising from different interventions and hence different distributions.

**Remark 1.** A version of Theorem 1 also holds without the faithfulness assumptions (A1) and (A2).

We define an *independence map* as a DAG  $D^*$  such that the observational distribution in (2) (or equivalently the distribution of  $X_{\text{obs}} \sim f_{\text{obs}}(x)dx$  in (6)) and all interventional distributions of  $X_{\text{int}}^{(i)} \sim f_{\text{int}}(x \mid \text{do}(X_{T^{(i)}} = U^{(i)}))dx$  in (6), for all  $T^{(i)}$ , can be generated by  $D^*$  and the corresponding intervention DAGs  $D_{T^{(i)}}^*$ . This is a generalization of an independence map for observational data (Pearl, 1988). A *minimum* independence map is an independence map having a minimum number of edges. A minimum independence map is typically not unique, and assuming faithfulness in (A1) and (A2), the set of all minimum independence maps equals the interventional Markov equivalence class  $\mathcal{E}_{\mathcal{I}_0}(D_0)$  with  $\mathcal{I} = \{T^{(i)}; i = 1, \dots, n_{\text{int}}\}$ .

Instead of (14) consider the estimator

$$\hat{D} = \arg \min_D -\ell_D(\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{T}, \mathbf{X}) + \frac{1}{2} \log(n) \dim(D),$$

where the optimization is over all DAGs  $D$ . The statement in Theorem 1 can then be replaced by:

$$\mathbb{P}[\hat{D} \text{ is a minimum independence map}] \rightarrow 1.$$

**Remark 2.** Although we have data sets with both observational and interventional data in mind, note that Theorem 6 only makes the assumption of a *conservative* family of intervention targets. In other words, consistent model selection is even possible with interventional data alone.

Let  $I \in \mathcal{I} \setminus \{\emptyset\}$  be an intervention target, and denote by  $n_I = |\{i; T^{(i)} = I, i = 1, \dots, n\}|$  the number of interventional data for this target. Assumption (A3) made in the theorem implies  $n_I \asymp n \rightarrow \infty$ . This might not be realistic in practice since there is often only one (or very few) interventional data point for each target  $I$ , i.e.,  $n_I = 1$  (or  $n_I$  is small). Without having a rigorous proof, the consistency result of Theorem 1 is expected to hold if the intervention value  $U^{(i)}$  is far away from zero, i.e., far away from the mean of  $X_{T^{(i)}}$ . The heuristics can be exemplified as follows.

**Example 1.** Consider a DAG  $D_0 = 1 \rightarrow 2$  with  $p = 2$  and corresponding observational distribution from the structural equation model

$$\begin{aligned} X_1 &\sim \mathcal{N}(0, \sigma_1^2), \\ X_2 &\leftarrow \beta X_1 + \varepsilon_2, \quad \varepsilon_2 \sim \mathcal{N}(0, \sigma_2^2). \end{aligned}$$



Then, the interventional distribution with target  $I = 1$  equals:

$$X_2 | \text{do}(X_1 = u) \sim \mathcal{N}(\beta u, \sigma_2^2), \quad (15)$$

whereas the marginal observational distribution is

$$X_2 \sim \mathcal{N}(0, \sigma_1^2 + \beta^2 \sigma_2^2). \quad (16)$$

Thus, if  $u \rightarrow \infty$ , the means of the distributions in (15) and (16) drift away from each other and one realization from the intervention in (15) would be sufficient such that with probability tending to 1 as  $u \rightarrow \infty$ , we could detect the difference from (one or many realizations of) the observational distribution in (16).

Alternatively, if  $u = 0$ , we could detect differences of the distributions in (15) and (16) in terms of their variances. But we would need many realizations from (15) and (16) to detect this difference with high probability.

Although obvious, we note that if the true DAG would be  $1 \leftarrow 2$ , the distribution in (15) and (16) would coincide (being equal to  $\mathcal{N}(0, \text{Var}(X_2))$ ). Therefore, when doing an intervention  $\text{do}(X_1 = u)$  and we see a difference in comparison to the marginal distribution of  $X_2$ , the true DAG must be  $1 \rightarrow 2$ .

We refer to empirical results in Section 4.2 which confirm good model selection properties if  $n_{\text{obs}}$  is large,  $n_I = 1$  but with intervention values  $U$  chosen sufficiently far away from zero.

### 3.3 Computation

The computation of the estimator in (14) is a highly non-trivial task. The main difficulty comes from the fact we have to optimize over all Markov equivalence classes. We can reformulate the optimization as follows:

$$\hat{B}, \{\hat{\sigma}_k^2\}_k = \arg \min_{B \in \mathbf{B}_{\text{DAG}}; \{\sigma_k^2\}_k} -\ell(B, \{\sigma_k^2\}_k; \mathcal{T}, \mathbf{X}) + \frac{1}{2} \log(n) \dim(B)$$

where  $-\ell(\cdot; \mathcal{T}, \mathbf{X})$  is the negative log-likelihood in the model (10), and  $\mathbf{B}_{\text{DAG}}$  is the space of matrices satisfying the constraint that they correspond to a DAG. This DAG-constraint causes the optimization to be highly non-convex. In view of this, the  $\ell_0$ -penalty is not adding major new computational challenges (and in fact allows for dynamic programming optimization, see below) while it enjoys nice statistical properties and leading to a score (value of the objective function) which is the same for all DAG members in an interventional Markov equivalence class.

Somewhat surprisingly, although the optimization problem in (14) is NP-hard (Chickering, 1996), dynamic programming can be used for exhaustive optimization (Silander and Myllymäki, 2006), roughly as long as  $p$  is less than say 20. For problems with larger dimension, the optimization in (14) can be pursued using greedy algorithms. Based on the idea from Chickering (2002a,b), one can use a greedy forward, backward and turning arrows algorithm which pursues each greedy step in the space of interventional Markov equivalence classes which is the much more appropriate space than the space of DAGs. An efficient implementation of such an algorithm, called Greedy Interventional Equivalent Search (GIES), is rigorously described in Hauser and Bühlmann (2012) where algorithmic properties, theoretical and empirical, are reported in detail. Although there is no guarantee that GIES converges to a global optimum, it seems very competitive and keeps up with dynamic programming for small-scale problems. An implementation of GIES is available in the R-package `pcalg` which is used throughout in Section 4.

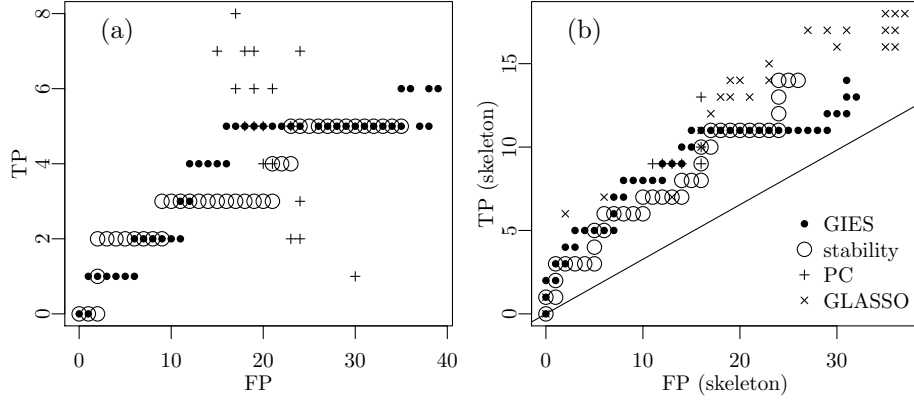


Figure 2: ROC plots of the models estimated from the Sachs data set, for directed edges (a) and the skeleton (b). In (a), GLASSO is missing since it does not yield a directed model; in (b), random guessing is shown by a solid line.

## 4 Empirical results

We evaluated  $\ell_0$ -penalized maximum likelihood estimation of interventional Markov equivalence classes as described in Section 3 on a real data set (Section 4.1) as well as on simulated data (Section 4.2).

### 4.1 Analysis of protein-signaling data

We analyzed the protein-signaling data set of Sachs et al. (2005). This data set contains 7466 measurements of the abundance of 11 phosphoproteins and phospholipids recorded under different experimental conditions in primary human immune system cells. The different experimental conditions are characterized by associated reagents that inhibit or activate signaling nodes, corresponding to interventions at different points of the protein-signaling network. Interventions mostly take place at more than one point, and the data set is purely interventional. However, some of the experimental perturbations affect receptor enzymes instead of (measured) signaling molecules. Since our statistical framework cannot cope with interventions at latent variables, we only considered 5846 out of the 7466 measurements which had an *identical* perturbation of the receptor enzymes. In this way, we model the system with perturbed receptor enzymes as “ground state”, defining its distribution of molecule abundances as observational.

While we can make the data set fit our interventional framework by the aforementioned reduction to 5846 data points, the linear-Gaussian assumption of our framework may not hold, even after a log-transformation of the measurements. Nevertheless, we fitted graphical models to the data set with different frequentist methods: GIES for the  $\ell_0$ -penalized MLE in (14) (see also Sections 3.3 and 4.2.1), the PC algorithm (Spirtes et al., 2000), the graphical LASSO (GLASSO, Friedman et al., 2007), and GIES combined with stability selection (Meinshausen and Bühlmann, 2010). We varied the tuning parameter of each algorithm: the number of steps (i.e., of edge additions, deletions or reversals) in GIES, the significance level  $\alpha$  in the PC algorithm, the penalty parameter  $\lambda$  in GLASSO, and the cut-off selection probability in stability selection applied for GIES. We compared the estimated models to the conventionally accepted model which serves as ground truth (Sachs et al., 2005); the resulting ROC plots, both with respect to edge directions (defining true and false positives in terms of the graphs’ adjacency matrices) and with respect to the skeleton alone, are depicted in Figure 4.1.

The overall performance of the estimation of the *skeleton* is comparable for all four algorithms (Figure 4.1(b)), even if two of them (PC and GLASSO) treat all data as identically distributed and disregard its interventional nature. Regarding edge directions (Figure 4.1(a)), however,

GIES (with or without stability selection) yields an improvement over the PC algorithm.

The Bayesian method of Cooper and Yoo (1999) used for model fitting by Sachs et al. (2005) is not directly comparable to the frequentist methods used here. In particular, the results from Sachs et al. (2005) are not easily reproducible due to choosing discretization levels and prior distribution. Their performance as measured by comparison to the ground truth is substantially better than all methods considered in this paper (15 true positives, 7 false positives in the convention of Figure 4.1(a)). Potential reasons are increased robustness due to discretization and specific tuning (which is legitimate in their context of extending and improving the conventional ground truth).

## 4.2 Simulations

We performed  $\ell_0$ -penalized maximum likelihood estimation as in (14) on interventional and observational data simulated from 4000 randomly drawn Gaussian causal models (see (6) or (10)) to illustrate the consistency result of Theorem 1.

### 4.2.1 Experimental Settings

We randomly drew DAGs whose skeleton has an expected vertex degree of 1.8, 1.9, 2.9 and 3.9 for  $p = 10, 20, 30$  and  $40$ , respectively. For every DAG  $D$ , we randomly generated a weight matrix  $B \in \mathbf{B}(D)$  and error variances  $\sigma_1^2, \dots, \sigma_p^2$  such that the corresponding observational covariance matrix

$$\Sigma = \text{Cov}(X_{\text{obs}}) = (\mathbb{I} - B)^{-1} \text{diag}(\sigma^2)(\mathbb{I} - B)^{-\text{T}}$$

had a diagonal of  $(1, \dots, 1)$ , meaning that each variable of the system had an observational marginal variance of 1. The procedure for generating Gaussian causal models of this form has been described in detail by Hauser and Bühlmann (2012).

We simulated data sets with a total sample size  $n = n_{\text{obs}} + n_{\text{int}}$  between 50 and 10'000. We performed single-vertex interventions at  $k$  randomly drawn vertices ( $k = 0.2p$ ,  $k = 0.5p$  and  $k = p$ ), drawing  $p/k$  samples under each intervention (that is, 5, 2 or 1 for the chosen values of  $k$ ). These settings ensure that we had only  $n_{\text{int}} = p$  interventional data points in each simulation, and that the majority of the data points were observational ones thus ( $n_{\text{obs}} = n - p$ ). This allowed us to verify our conjecture following Theorem 1 that few interventional samples are sufficient for consistent estimation of interventional Markov equivalence classes (or, equivalently, interventional essential graphs) as long as the intervention levels, the expectation values of the intervention variables  $U$ , are large enough. In our simulations, we chose expectation values  $\mu_{U_j}$  between 1 and 50 and variances of  $\tau^2 = (0.2)^2$  for the intervention variables. Note that because of the chosen normalization  $\Sigma_{ii} = 1$ , the expectation values  $\mu_{U_j}$  can be thought of as being indicated in units of observational standard deviations.

To sum up, for each of the 4000 randomly generated Gaussian causal models, we simulated 144 data sets with observational and interventional data, namely one data set for each combination of the following experimental parameters:

- $n \in \{50, 100, 200, 500, 1000, 2000, 5000, 10000\}$ ;  $n_{\text{int}} = p$ ,  $n_{\text{obs}} = n - p$ ;
- $k \in \{0.2p, 0.5p, p\}$ ;
- $\mu_{U_j} \in \{1, 2, 5, 10, 20, 50\}$ .

We learned the structure of the underlying causal model from the simulated data sets using the BIC score as described in Section 3.3. We used the two causal inference algorithms mentioned in Section 3.3:

- an adaptation of the dynamic programming approach of Silander and Myllymäki (2006) to interventional data which will be abbreviated as SiMY in the following. This algorithm

guarantees to find the global minimizer of the BIC in (14); because of its exponential complexity, it is only applicable for models with no more than 20 variables though.

- the Greedy Interventional Equivalence Search (GIES) of Hauser and Bühlmann (2012). This algorithm *greedily* optimizes the BIC score by traversing the search space of interventional Markov equivalence classes through operations corresponding to edge additions, deletions, or reversals in the space of DAGs. The algorithm does not *guarantee* to find the optimum of the BIC, but it was empirically shown for graphs with up to  $p = 20$  nodes to have a performance comparable to that of SiMY (Hauser and Bühlmann, 2012) while having polynomial runtime in the average case.

We assessed the quality of the estimated causal models with the structural Hamming distance SHD (Tsamardinos et al., 2006; we use the slightly adapted version of Kalisch and Bühlmann, 2007). This quantity is a metric on the space of graphs. The SHD between two graphs  $G$  and  $\hat{G}$  is the sum of false positives and false negatives of the skeleton and wrongly oriented edges. Formally, if the graphs  $G$  and  $\hat{G}$  have adjacency matrices  $A$  and  $\hat{A}$ , respectively, the SHD between  $G$  and  $\hat{G}$  is defined as

$$\text{SHD}(G, \hat{G}) := \sum_{1 \leq i < j \leq p} (1 - \mathbb{1}_{\{(A_{ij} = \hat{A}_{ij}) \wedge (A_{ji} = \hat{A}_{ji})\}}) .$$

#### 4.2.2 Results

Figure 3 shows the SHD between estimated and true interventional essential graph as a function of the total sample size  $n$ . Results for different numbers of intervention targets showed similar characteristics (not shown). The plots illustrate the consistency of the BIC, the main result of Theorem 1. As expected, convergence to the true equivalence class is faster the larger the intervention values (controlled by  $\mu_U$ ) are. Note, however, that the simulation setting does not fully match the limit setting of the theorem: while the theoretical result asks for the sample sizes  $n_I$  of *all* interventions  $I \in \mathcal{I}$  to grow in the order  $O(n)$ , we always have  $p$  interventional data points in our case while only the number of observational data points is growing. In the setting with  $n_I \asymp n$ , Hauser and Bühlmann (2012) have already empirically shown the performance of GIES as well as SiMY.

Figure 4 supports our conjecture stated after Theorem 1: even with few interventional data points (a total of  $p$  in our case, compared to  $n - p \gg p$  for  $n = 1000$ ), the estimates of the causal models are substantially improved by only increasing the mean intervention values  $\mu_U$ . However, for  $p = 10$ , this effect is not clearly visible.

## 5 Conclusions

We have proposed a likelihood framework for joint modeling of Gaussian interventional and observational data. Such kind of data arises in many applications, notably in biology with measurements of wild-type individuals and modifications arising from interventional knock-outs of some genes. Our likelihood approach has various interesting aspects which we summarize as follows. The parameters in the model are the observational directed acyclic graph (DAG)  $D$  and the corresponding edge weights  $B$  and error variances  $\{\sigma_i^2\}_i$  (or instead of  $B$  and  $\{\sigma_i^2\}_i$  the corresponding covariance matrix of a Gaussian distribution). These parameters are global in the sense that every intervention distribution is determined by these parameters via the do-calculus: in particular, this implies that only one or a few data per intervention suffice for reasonably accurate estimation since the corresponding distributions are all linked to the global parameters.

We show here that the BIC is consistent for estimating the corresponding interventional Markov equivalence class. The proof is rather involved since the various intervention distributions are not identical and do not easily fit into a standard setting. The interventional Markov

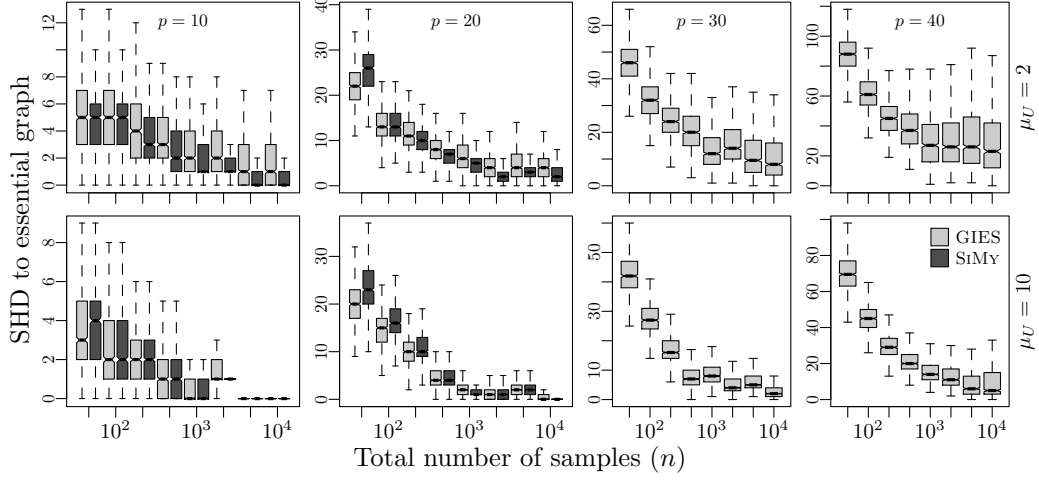


Figure 3: SHD between estimated and true interventional essential graph as a function of the sample size  $n$  for different numbers of variables  $p$ . In each simulation,  $p$  interventional data points were used, 2 replicates for  $p/2$  single-vertex intervention targets. Interventions were performed with an expectation value of  $\mu_U = 2$  (upper row) and  $\mu_U = 10$  (lower row), respectively.

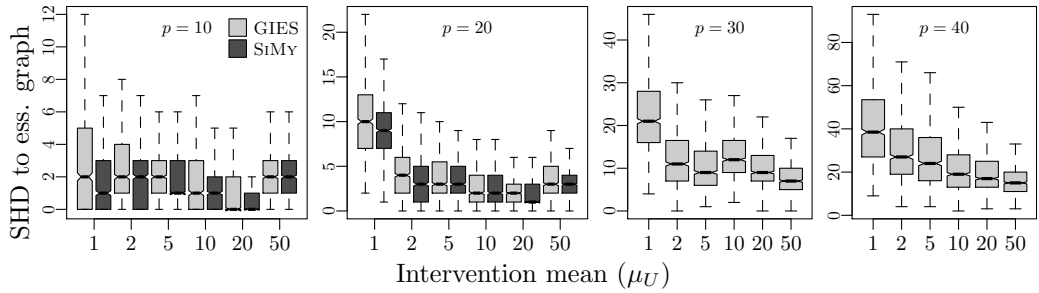


Figure 4: SHD between estimated and true interventional essential graph as a function of the intervention mean  $\mu_U$ . Results for simulations with a total sample size of  $n = 1000$ , of which  $p$  data points originate from interventions at 20% of the vertices.

equivalence class is an interesting and realistic target: it is smaller than the standard observational Markov equivalence class and it leads to a higher degree of identifiability when intervening at several variables. This has direct implications for tighter bounds for inferring causal effects (Maathuis et al., 2009).

Besides the methodological development and theoretical derivations, we present empirical results for real and simulated data.

## 6 Derivations and proofs

This section contains all proofs left out in earlier sections, namely the derivation of the maximum likelihood estimator for a given DAG (Section 6.1, proving results of Section 2.2), and the proof of the consistency result for model selection (Section 6.3 proving Theorem 1).

### 6.1 Explicit form of maximum likelihood estimator when DAG is known

Gaussian densities form an exponential family. The joint density of Gaussian random variables with expectation  $\mu$  and covariance  $\Sigma$  can be written as

$$f_{\mathcal{N}}(x; K, \nu) = (2\pi)^{-1/2} \exp \left[ \left\langle -\frac{1}{2}xx^T, K \right\rangle_{\mathcal{S}^p} + \langle x, \nu \rangle_{\mathbb{R}^p} - \frac{1}{2}(\nu^T K^{-1} \nu - \log \det K) \right] , \quad (17)$$

where the inverse covariance matrix or precision matrix  $K := \Sigma^{-1}$  and the transformed expectation value  $\nu := K\mu$  form the natural parameters. In (17),  $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$  stands for the canonical inner product on  $\mathbb{R}^p$ , and  $\langle \cdot, \cdot \rangle_{\mathcal{S}^p}$  denotes the inner product  $\langle A, B \rangle_{\mathcal{S}^p} := \text{tr}(AB)$  on the vector space  $\mathcal{S}^p$  of symmetric  $p \times p$  matrices.

The canonical form (17) of the exponential family of Gaussian distributions eases calculations with the interventional distributions (11), especially for our goal to derive a maximum likelihood estimator for a causal model with interventional data originating from *different* interventions. We hence start by calculating the natural parameters for the interventional distribution (11). To simplify later calculations, we use the inverse error variances  $\gamma_k := \sigma_k^{-2}$  to parameterize a Gaussian causal model from here on, together with the vector notation  $\gamma := (\gamma_1, \dots, \gamma_p)$ .

**Lemma 2** *Let  $\mu^{(I)}$  and  $\Sigma^{(I)}$  be the expectation and covariance of the interventional distribution (11), respectively. Then the following identities hold:*

$$\begin{aligned} K^{(I)} &:= (\Sigma^{(I)})^{-1} = (\mathbb{I} - B)^T R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B) + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)} , \\ \nu^{(I)} &:= K^{(I)} \mu^{(I)} = Q^{(I)T} \tilde{K}^{(I)} \mu_{U_I} , \\ \nu^{(I)T} (K^{(I)})^{-1} \nu^{(I)} &= \mu_{U_I}^T \tilde{K}^{(I)} \mu_{U_I} , \\ \log \det K^{(I)} &= \sum_{k \notin I} \log \gamma_k + \log \det \tilde{K}^{(I)} . \end{aligned}$$

We make use here of the notation  $\tilde{K}^{(I)} := (\tilde{\Sigma}^{(I)})^{-1}$ ;  $\tilde{\Sigma}^{(I)} := \text{diag}(\tau_I^2)$  is the covariance matrix of the intervention variable  $U_I$ .

**Proof** To prove the formulae, we use the following identities of the auxiliary matrices (12):

$$\begin{aligned} P^{(I)} P^{(I)T} &= \mathbb{I} & P^{(I)} Q^{(I)T} &= 0 & Q^{(I)T} Q^{(I)} &= \mathbb{I} - R^{(I)} \\ Q^{(I)} Q^{(I)T} &= \mathbb{I} & Q^{(I)} P^{(I)T} &= 0 & R^{(I)} R^{(I)} &= R^{(I)} \end{aligned} \quad (18)$$

To verify the claimed formula for the precision matrix  $K^{(I)}$ , it can be easily checked using the identities (18) that

$$\left[ R^{(I)} \text{diag}(\sigma^2) R^{(I)} + Q^{(I)T} \tilde{\Sigma}^{(I)} Q^{(I)} \right]^{-1} = R^{(I)} \text{diag}(\gamma) R^{(I)} + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)} . \quad (19)$$

We then find

$$\begin{aligned} K^{(I)} &\stackrel{(11)}{=} \left( \mathbb{I} - R^{(I)B} \right)^T \left[ R^{(I)} \text{diag}(\gamma) R^{(I)} + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)} \right] \left( \mathbb{I} - R^{(I)B} \right) \\ &= (\mathbb{I} - B)^T R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B) + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)}, \end{aligned}$$

where we again use several of the identities (18) in the last step.

By making use of equations (11) and (19) again, we can calculate the transformed expectation:

$$\begin{aligned} \nu^{(I)} &= \left( \mathbb{I} - R^{(I)} R \right)^T \left[ R^{(I)} \text{diag}(\gamma) R^{(I)} + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)} \right] Q^{(I)T} \mu_{U_I} \\ &= Q^{(I)T} \tilde{K}^{(I)} \mu_{U_I}; \end{aligned}$$

the last step is again a consequence of the identities (18).

For the next formula, we use the fact that  $B$  is a nilpotent matrix; it is not hard to see that every matrix satisfying the DAG constraint actually is nilpotent. Therefore the inverse of  $(\mathbb{I} - R^{(I)} B)$  can be calculated as  $(\mathbb{I} - R^{(I)} B)^{-1} = \sum_{k=0}^{p-1} (R^{(I)} B)^k$ . Together with the identities (18) and the representation of  $\mu^{(I)}$  in (11), we conclude that

$$Q^{(I)} \mu^{(I)} = \sum_{k=0}^{p-1} Q^{(I)} \left( R^{(I)} B \right)^k Q^{(I)T} \mu_{U_I} = \mu_{U_I}.$$

It follows that

$$\nu^{(I)T} \left( K^{(I)} \right)^{-1} \nu^{(I)} = \mu^{(I)T} \nu^{(I)} = \mu^{(I)T} Q^{(I)T} \tilde{K}^{(I)} \mu_{U_I} = \mu_{U_I}^T \tilde{K}^{(I)} \mu_{U_I},$$

where we used the formula for  $\nu^{(I)}$  already proven before.

To calculate the determinant of  $K^{(I)}$  finally, note that there is a permutation matrix  $P$  such that

$$P \left[ R^{(I)} \text{diag}(\gamma) R^{(I)} + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)} \right] P^T$$

is a block matrix. Hence

$$\det K^{(I)} = \det \tilde{K}^{(I)} \cdot \prod_{k \notin I} \gamma_k$$

or  $\log \det K^{(I)} = \sum_{k \notin I} \log \gamma_k + \log \det \tilde{K}^{(I)}$ , which completes the proof.  $\square$

Up to now, we have only considered a single interventional distribution. In the next lemma, we provide a formula for the likelihood of an interventional dataset originating from *multiple* intervention targets as defined in (9). In the following, we simplify notation by unifying observational and interventional data point in a common framework. For this aim, we reuse the convention at the end of Section 2.2 and consider the *entire* data set  $(X^{(i)})_{i=1}^n$ ,  $n = n_{\text{obs}} + n_{\text{int}}$ , of all observational and interventional data points. To make notation short, we denote the complete data set by the matrix  $\mathbf{X}$ , having the rows  $X^{(1)}, \dots, X^{(n)}$ , and the list of intervention targets  $T^{(1)}, \dots, T^{(n)}$  by  $\mathcal{T}$ . Recall that an observational data point  $X^{(i)}$  is marked by the empty target  $T^{(i)} = \emptyset$ .

**Lemma 3** *Let  $(\mathcal{T}, \mathbf{X})$  be an interventional dataset as defined above, produced by a Gaussian causal model with structure  $D$ . Moreover, let  $B \in \mathbf{B}(D)$  be a weight matrix and  $\gamma \in \mathbb{R}_{>0}^p$  a vector of inverse error variances. Denote by  $n^{(I)} := |\{i \mid T^{(i)} = I\}|$  and  $S^{(I)} := \frac{1}{n^{(I)}} \sum_{i: T^{(i)} = I} X^{(i)} X^{(i)T}$*

(empirical covariance matrix for intervention  $I \in \mathcal{I}$ ). Then the log-likelihood of  $(\mathcal{T}, \mathbf{X})$  given parameters  $B$  and  $\gamma$  is

$$\begin{aligned}\ell_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= -\frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \text{tr} \left( S^{(I)} K^{(I)} \right) + \frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \log \det K^{(I)} + C \\ &= -\frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \text{tr} \left[ S^{(I)} (\mathbb{I} - B)^T R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B) \right] + \frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \sum_{k \notin I} \log \gamma_k + C',\end{aligned}$$

where  $C$  and  $C'$  are constants given by the dataset  $(\mathcal{T}, \mathbf{X})$  that do not depend on the model parameters  $B$  and  $\gamma$ .

Note that in the case of purely observational data (that is, if  $T^{(i)} = \emptyset$  for all  $i$ ), this result reproduces the classical log-likelihood (see, for example, Banerjee et al., 2008)

$$2\ell_D(B, \gamma; (\emptyset)_{i=1}^n, \mathbf{X}) = n(\log \det K - \text{tr}(SK)) + C.$$

**Proof** The likelihood of the entire data set is the product of the sample likelihoods (11):

$$\begin{aligned}\ell_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= \sum_{i=1}^n \log f \left( X^{(i)} \mid \text{do}(X_{T^{(i)}}^{(i)} = U_{T^{(i)}}^{(i)}) \right) \\ &\stackrel{(11)}{=} \sum_{i=1}^n \log f_{\mathcal{N}} \left( X^{(i)}; K^{(T^{(i)})}, \nu^{(T^{(i)})} \right) \\ &\stackrel{(17)}{=} -\frac{1}{2} \sum_{i=1}^n \text{tr} \left( X^{(i)} X^{(i)T} K^{(T^{(i)})} \right) + \frac{1}{2} \sum_{i=1}^n \log \det K^{(T^{(i)})} + C \\ &= -\frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \text{tr} \left( S^{(I)} K^{(I)} \right) + \frac{1}{2} \sum_{I \in \mathcal{I}} n^{(I)} \log \det K^{(I)} + C.\end{aligned}$$

In the calculations above,  $C$  stands for a constant that is independent of the model parameters  $B$  and  $\gamma$  (note that, by Lemma 2, the remaining terms from Equation (17) are independent of model parameters).

The second line of the lemma follows easily from the first one by applying the identities given in Lemma 2.  $\square$

The following lemma shows that the log-likelihood derived before is *decomposable* (Chickering, 2002b) in the sense that it can be written as a sum of terms that only depend on a vertex and its parents.

**Lemma 4** Using the definitions  $n^{(-k)} := \sum_{I \in \mathcal{I}: k \notin I} n^{(I)}$  and  $S^{(-k)} := \sum_{I \in \mathcal{I}: k \notin I} \frac{n^{(I)}}{n^{(-k)}} S^{(I)}$ , the log-likelihood of Lemma 3 can be decomposed as follows:

$$\begin{aligned}\ell_D(B, \gamma; \mathcal{T}, \mathbf{X}) &= \sum_{k=1}^p \ell_k(B_{k\bullet}, \gamma_k; \mathcal{T}, \mathbf{X}) + C, \\ \ell_k(B_{k\bullet}, \gamma_k; \mathcal{T}, \mathbf{X}) &= -\frac{1}{2} n^{(-k)} \left[ \gamma_k (\mathbb{I} - B)_{k\bullet} S^{(-k)} ((\mathbb{I} - B)_{k\bullet})^T - \log \gamma_k \right],\end{aligned}$$

where  $C$  is a constant that does not depend on the parameters  $\gamma$  and  $B$ . The calculation of the partial likelihoods  $\ell_k$  only involves data measured at vertex  $k$  and its parents  $\text{pa}(k)$ .

**Proof** The decomposition of the second summand in Lemma 3 is easy to verify:

$$\sum_{I \in \mathcal{I}} n^{(I)} \sum_{k \notin I} \log \gamma_k = \sum_{i=1}^n \sum_{k \notin T^{(i)}} \log \gamma_k = \sum_{k=1}^p \sum_{i: k \notin T^{(i)}} \log \gamma_k = \sum_{k=1}^p n^{(-k)} \log \gamma_k.$$



The decomposition of the first summand makes use of the fact that  $\text{tr}(AB) = \text{tr}(BA)$  for any matrices  $A$  and  $B$  for which  $AB$  and  $BA$  are defined:

$$\begin{aligned}
& \sum_{I \in \mathcal{I}} n^{(I)} \text{tr} \left[ S^{(I)} (\mathbb{I} - B)^T R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B) \right] \\
&= \sum_{i=1}^n \text{tr} \left[ R^{(T(i))} \text{diag}(\gamma) R^{(T(i))} (\mathbb{I} - B) X^{(i)} X^{(i)T} (\mathbb{I} - B)^T \right] \\
&= \sum_{i=1}^n \sum_{k \notin T(i)} \gamma_k (\mathbb{I} - B)_{k\bullet} X^{(i)} X^{(i)T} ((\mathbb{I} - B)_{k\bullet})^T \\
&= \sum_{k=1}^p n^{(-k)} \gamma_k (\mathbb{I} - B)_{k\bullet} S^{(-k)} ((\mathbb{I} - B)_{k\bullet})^T.
\end{aligned}$$

The  $k_{\text{th}}$  column of  $\mathbb{I} - B$ ,  $(\mathbb{I} - B)_{k\bullet}$  only has entries at indices  $\{k\} \cup \text{pa}(k)$ , so the calculation only includes rows and columns of the empirical covariance matrix with those indices and hence only uses data from vertex  $k$  and its parents.  $\square$

Lemma 4 shows that, for a fixed DAG  $D$ , the maximum likelihood estimates for the weight matrix and the error variances can be calculated “locally”, that is only involving data of single vertices and their parents.

**Lemma 5** *For a fixed DAG  $D$  and given data, the maximum likelihood estimate for its parameters  $\sigma$  and  $B$  are*

$$\hat{B}_{k, \text{pa}(k)} = S_{k, \text{pa}(k)}^{(-k)} \left( S_{\text{pa}(k), \text{pa}(k)}^{(-k)} \right)^{-1}, \quad \hat{\sigma}_k^2 = (\mathbb{I} - \hat{B})_{k\bullet} S^{(-k)} \left( (\mathbb{I} - \hat{B})_{k\bullet} \right)^T,$$

The maximum partial likelihoods are

$$\begin{aligned}
\sup_{B_{k\bullet}, \gamma_k} \ell_k(B_{k\bullet}, \gamma_k; \mathcal{T}, \mathbf{X}) &= -\frac{1}{2} n^{(-k)} (1 + \log \hat{\sigma}_k^2) \\
&= -\frac{1}{2} n^{(-k)} \left\{ 1 + \log \left[ S_{kk}^{(-k)} - S_{k, \text{pa}(k)}^{(-k)} \left( S_{\text{pa}(k), \text{pa}(k)}^{(-k)} \right)^{-1} S_{\text{pa}(k), k}^{(-k)} \right] \right\}
\end{aligned}$$

**Proof** The maximum likelihood estimate must be a root of the derivative of the likelihood. From Lemma 4, we see that  $\frac{\partial}{\partial B_{ki}} \ell = \frac{\partial}{\partial B_{ki}} \ell_k$  for  $i = 1, \dots, p$ . This partial derivative is

$$\frac{\partial}{\partial B_{ki}} \ell_k(B_{k\bullet}, \gamma_k; \mathcal{T}, \mathbf{X}) \propto (\mathbb{I} - B)_{k\bullet} S_i^{(-k)} = S_{ki}^{(-k)} - B_{k\bullet} S_i^{(-k)}. \quad (20)$$

For a fixed  $k$ ,  $B_{k\bullet}$  has one non-zero entry for every parent of  $k$  in the DAG  $D$ . For those entries, we get the system of linear equations

$$B_{k, \text{pa}(k)} S_{\text{pa}(k), i}^{(-k)} = S_{ki}^{(-k)}, \quad \forall i \in \text{pa}(k),$$

by setting the partial derivatives (20) to zero. In matrix notation, this reads

$$B_{k, \text{pa}(k)} S_{\text{pa}(k), \text{pa}(k)}^{(-k)} = S_{k, \text{pa}(k)}^{(-k)}$$

and has the solution

$$\hat{B}_{k, \text{pa}(k)} = S_{k, \text{pa}(k)}^{(-k)} \left( S_{\text{pa}(k), \text{pa}(k)}^{(-k)} \right)^{-1};$$

note that  $S_{k, \text{pa}(k)}^{(-k)}$  is invertible almost surely if  $n^{(-k)} > |\text{pa}(k)|$ .

The derivative with respect to the error variances is

$$\frac{\partial}{\partial \gamma_k} \ell_k(B_{k\bullet}, \gamma_k) \propto (\mathbb{I} - B)_{k\bullet} S^{(-k)} ((\mathbb{I} - B)_{k\bullet})^T - \frac{1}{\gamma_k}$$

and has the inverse root

$$\begin{aligned} \frac{1}{\hat{\gamma}_k} &= \hat{\sigma}_k^2 = (\mathbb{I} - \hat{B})_{k\bullet} S^{(-k)} ((\mathbb{I} - \hat{B})_{k\bullet})^T \\ &= S_{kk}^{(-k)} - S_{k, \text{pa}(k)}^{(-k)} \left( S_{\text{pa}(k), \text{pa}(k)}^{(-k)} \right)^{-1} S_{\text{pa}(k), k}^{(-k)}. \end{aligned}$$

By plugging this into the formula of Lemma 4, we immediately find the formula for the supremum of the partial likelihoods.  $\square$

## 6.2 Definition of interventional Markov equivalence class

The observational Markov equivalence class of a DAG can be described as follows. For a DAG  $D$ , denote by  $\mathcal{M}(D) = \{f; f \text{ Markov with respect to } D\}$  all distributions which are Markov with respect to  $D$ . Thereby, the Markovian property is meant to be the factorization property as in (3), and we denote by  $f$  the density of the  $p$ -dimensional Gaussian distribution. Two DAGs  $D \sim D'$  are Markov equivalent, if and only if  $\mathcal{M}(D) = \mathcal{M}(D')$ . The observational equivalence class of a DAG  $D$  is then denoted by  $\mathcal{E}(D)$  which can be represented as an essential graph which is a chain graph with directed and undirected edges (Andersson et al., 1997).

For the interventional Markov equivalence class, we proceed as follows. For a DAG  $D$ , consider the corresponding intervention DAG  $D_I$  where we remove all edges which point from  $\text{pa}(I)$  to  $I$ . Furthermore, consider a family of intervention targets  $\mathcal{I}$  and corresponding tuples of densities  $(f_I)_{I \in \mathcal{I}}$ , where each element corresponds to an intervention target  $I \in \mathcal{I}$ . Let

$$\begin{aligned} \mathcal{M}_{\mathcal{I}}(D) &= \{(f_I)_{I \in \mathcal{I}}; \quad \forall I \in \mathcal{I}: f_I \in \mathcal{M}(D_I), \text{ and} \\ &\quad \forall I, J \in \mathcal{I}, \forall i \notin I \cup J: f_I(x_i | x_{\text{pa}_D(i)}) = f_J(x_i | x_{\text{pa}_D(i)})\}. \end{aligned}$$

Two DAGs  $D$  and  $D'$  are interventionally Markov equivalent with respect to the family of targets  $\mathcal{I}$  (notation:  $D \sim_{\mathcal{I}} D'$ ) if and only if  $\mathcal{M}_{\mathcal{I}}(D) = \mathcal{M}_{\mathcal{I}}(D')$  (Hauser and Bühlmann, 2012). For a DAG  $D$ , the interventional Markov equivalence class with respect to  $\mathcal{I}$  (or  $\mathcal{I}$ -Markov equivalence class) is denoted by  $[D]_{\mathcal{I}}$  which, as in the observational case, can be characterized by an essential graph  $\mathcal{E}_{\mathcal{I}}(D)$  (Hauser and Bühlmann, 2012). For  $\mathcal{I} = \emptyset$ , the definition coincides with the observational Markov equivalence class above. Although the definition of interventional Markov equivalence is somewhat cumbersome, the defined object indeed represents the DAGs which are equivalent and non-distinguishable from the interventional distributions (and if  $\mathcal{I}$  also contains the  $\emptyset$ -target, from observational and interventional distributions). In other words, assuming faithfulness as in (2), the  $\mathcal{I}$  interventional Markov equivalence is identifiable from the distributions.

## 6.3 Proof of Theorem 1

In the previous section, we calculated the maximum of the likelihood of causal models given a set of interventional data  $(\mathcal{T}, \mathbf{X})$ . For model selection, that is, estimating the causal model that produced a given dataset, the model complexity has to be penalized to avoid overfitting. For large interventional (and potentially observational) samples, it stands to reason to choose the complexity penalty of the Bayesian information criterion (BIC).

The maximization of the BIC of a growing sequence of i.i.d. data is known to lead to consistent model selection from a set of curved exponential models (Haughton, 1988).

**Definition 1 (Curved exponential model; Haughton, 1988)** Let  $\mathcal{P} = \{f(x; \theta) = h(x) \exp[\langle T(x), \theta \rangle - b(\theta)] \mid \theta \in \Theta\}$  be an exponential family with natural parameter space  $\Theta \subset \mathbb{R}^k$ . A curved exponential model is a set of parameters of the form  $M \cap \Theta$ , where  $M$  is a smooth connected manifold embedded in  $\mathbb{R}^k$ .

Suppose  $(X^{(i)})_{i=1}^n$  is a sequence of i.i.d. realizations from a density in the exponential family of Definition 1, and let  $M \cap \Theta$  be an curved exponential model in that family. The *Bayesian information criterion* or *BIC* of  $M \cap \Theta$  is then defined as

$$\begin{aligned} S(M; \mathbf{X}) &:= \sup_{\theta \in M \cap \Theta} \log \prod_{i=1}^n f(X^{(i)}; \theta) - \frac{1}{2} \dim(M) \log n \\ &= n \sup_{\theta \in M \cap \Theta} (\langle \bar{T}_n, \theta \rangle - b(\theta)) - \frac{1}{2} \dim(M) \log n, \end{aligned} \quad (21)$$

where  $\bar{T}_n$  stands for the mean statistic  $\bar{T}_n := \frac{1}{n} \sum_{i=1}^n T(X^{(i)})$  and  $\mathbf{X}$  is the data matrix having the samples  $X^{(i)}$  as rows.

**Theorem 6 (Consistency of the BIC; Haughton, 1988)** Let  $M_1 \cap \Theta, M_2 \cap \Theta, \dots$  be a finite set of curved exponential models in the natural parameter space  $\Theta$  of an exponential family as in Definition 1 with the following property: for each  $i \neq j$ , if a point in  $\bar{M}_i$  is in  $M_j \cap \overset{\circ}{\Theta}$ , then it is in  $M_i$ .

Assume  $\theta \in \overset{\circ}{\Theta}$  and let  $M_i$  and  $M_j$  be two different curved exponential models. If  $\theta \in M_i \setminus M_j$ , or if  $\theta \in M_i \cap M_j$  with  $\dim(M_i) < \dim(M_j)$ , then

$$\lim_{n \rightarrow \infty} P_\theta[S(M_i; \mathbf{X}) > S(M_j; \mathbf{X})] = 1.$$

As we explained in Section 3.2, we regard the intervention targets  $T^{(1)}, \dots, T^{(n)}$  as a random sequence, taking a “value”  $I \in \mathcal{I}$  with probability  $w_I$  (assumption (A3) of Section 3.2). With this assumption, we can treat the complete data set  $(T^{(i)}, X^{(i)})_{i=1}^n$  as i.i.d. realizations of random variables  $(T, X) \in \mathcal{I} \times \mathbb{R}^p$ . Expressed in this notation, we have shown in Section 6.1 that the conditional densities  $f(x \mid T = I) = f_{\text{int}}(x \mid \text{do}(X_I = U_I))$  belong to an exponential family. In the next proposition, we show that also the joint density of  $(T, X)$  belongs to an exponential family.

**Proposition 7** Consider a set of random variables  $(X, Y) \in \mathbb{R}^p \times \{1, \dots, J\}$  with  $P[Y = j] = w_j$ ,  $1 \leq j \leq J$ , and  $(X \mid Y = j) \sim f(\cdot; \theta_j)$ , where  $f(x; \theta)$  is a density from an exponential family:

$$f(x; \theta) = h(x) \exp[\langle T(x), \theta \rangle - b(\theta)].$$

Then the joint density of  $X$  and  $Y$  is also an element of an exponential family, namely

$$f(x, y; \boldsymbol{\theta}, \eta) = h(x) \exp \left[ \left\langle S(x, y), \begin{pmatrix} \boldsymbol{\theta} \\ \eta \end{pmatrix} \right\rangle - a(\boldsymbol{\theta}, \eta) \right].$$

The natural parameters are given by  $\boldsymbol{\theta} = (\theta_1^T, \dots, \theta_J^T)^T$  and  $\eta = (\eta_1, \dots, \eta_{J-1})^T$  with  $\eta_j = \log \frac{w_j}{w_J} - b(\theta_j) + b(\theta_J)$ . The sufficient statistic  $S$  and the log-partition function  $a$  are given by

$$\begin{aligned} S(x, y) &= (\delta_{y,1} T(x)^T, \dots, \delta_{y,J} T(x)^T, \delta_{y,1}, \dots, \delta_{y,J-1})^T, \\ a(\boldsymbol{\theta}, \eta) &= b(\theta_J) + \log \left[ 1 + \sum_{j=1}^{J-1} \exp(\eta_j + b(\theta_j) - b(\theta_J)) \right]. \end{aligned}$$

**Proof** A straight-forward calculation yields to the claimed result:

$$\begin{aligned}
f(x, y; \boldsymbol{\theta}, \eta) &= w_y f(x; \theta_y) \\
&= h(x) \exp[\langle T(x), \theta_y \rangle - b(\theta_y) + \log w_y] \\
&= h(x) \exp \left[ \sum_{j=1}^J \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^J \delta_{y,j} (\log w_j - b(\theta_j)) \right] \\
&= h(x) \exp \left[ \sum_{j=1}^J \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^{J-1} \delta_{y,j} \left( \log \frac{w_j}{w_J} - b(\theta_j) + b(\theta_J) \right) \right. \\
&\quad \left. + \left( 1 - \sum_{j=1}^{J-1} \delta_{j,y} \right) (\log w_J - b(\theta_J)) + \sum_{j=1}^{J-1} \delta_{j,y} (\log w_J - b(\theta_J)) \right] \\
&= h(x) \exp \left[ \sum_{j=1}^J \langle \delta_{y,j} T(x), \theta_j \rangle - \sum_{j=1}^{J-1} \delta_{y,j} \left( \log \frac{w_j}{w_J} - b(\theta_j) + b(\theta_J) \right) \right. \\
&\quad \left. + \log w_J - b(\theta_J) \right] \\
&= h(x) \exp \left[ \left\langle S(x, y), \begin{pmatrix} \boldsymbol{\theta} \\ \eta \end{pmatrix} \right\rangle + \log w_J - b(\theta_J) \right]
\end{aligned}$$

with the definitions of  $S(x, y)$ ,  $\boldsymbol{\theta}$  and  $\eta$  from above.

To finish the calculation, we need to express  $w_J$  as a function of  $\boldsymbol{\theta}$  and  $\eta$ : since

$$w_J = 1 - \sum_{j=1}^{J-1} w_j = 1 - \sum_{j=1}^{J-1} \exp[\eta_j + b(\theta_j) - b(\theta_J)] w_J,$$

we find

$$w_J = \left[ 1 + \sum_{j=1}^{J-1} \exp(\eta_j + b(\theta_j) - b(\theta_J)) \right]^{-1},$$

what immediately yields the claimed log-partition function  $a(\boldsymbol{\theta}, \eta)$ .  $\square$

In order to prove the consistency of the BIC for causal model selection under interventions in the limit of large interventional samples, we must show that the models described by different DAGs fit the prerequisites of Theorem 6.

We have already seen that a single Gaussian interventional density (11) is a representative of an exponential family with natural parameters  $K^{(I)}$  and  $\nu^{(I)}$  living in  $\mathcal{S}^p$  and  $\mathbb{R}^p$ , respectively (see (17)). By Proposition 7, we conclude that the natural parameter space for the complete family of interventions is

$$\underbrace{(\mathcal{S}_{>0}^p)^J}_{=: \mathcal{S}} \times \underbrace{(\mathbb{R}^p)^J}_{=: \mathcal{V}} \times \underbrace{\mathbb{R}^{J-1}}_{=: \mathcal{W}},$$

where we write  $J := |\mathcal{I}|$ . We have already seen that the interventional densities are determined by *model parameters* and *experimental parameters*; the model parameters are  $B \in \mathbf{B}(D)$  and  $\gamma \in \mathbb{R}_{>0}^p$ . Therefore the sets of natural parameters corresponding to different models are parameterized by functions

$$\Phi_D^{\mathcal{I}} : \mathbf{B}(D) \times \mathbb{R}_{>0}^p \rightarrow \mathcal{S} \times \mathcal{V} \times \mathcal{W}.$$

Before showing that the images of those maps form indeed a set of embedded manifolds in  $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$  satisfying the prerequisites of Theorem 6, we sum up our notation from above and from Lemma 2.

**Definition 2** Let  $D$  be a DAG. Furthermore, let  $\mathcal{I}$  be a conservative family of intervention targets, and  $T \in \mathcal{I}$  arbitrary. Then we define

$$\begin{aligned} \Phi_D^{\mathcal{I}} : \mathbf{B}(D) \times \mathbb{R}_{>0}^p &\rightarrow \mathcal{S} \times \mathcal{V} \times \mathcal{W}, \\ (B, \gamma) &\mapsto \left( (K^{(I)}(B, \gamma))_{I \in \mathcal{I}}, (\nu^{(I)})_{I \in \mathcal{I}}, (\eta^{(I)})_{I \in \mathcal{I} \setminus \{T\}} \right) \end{aligned}$$

with

$$\begin{aligned} K^{(I)}(B, \gamma) &= (\mathbb{I} - B)^T R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B)^T + Q^{(I)T} \tilde{K}^{(I)} Q^{(I)}, \\ \nu^{(I)} &= Q^{(I)T} \tilde{K}^{(I)} \tilde{\mu}^{(I)}, \\ \eta^{(I)} &= \log \frac{\tilde{w}_I}{\tilde{w}_T} - b[K^{(I)}(B, \gamma), \nu^{(I)}] + b[K^{(T)}(B, \gamma), \nu^{(T)}], \\ b(K, \nu) &= \frac{1}{2}(\nu^T K^{-1} \nu - \log \det K). \end{aligned}$$

Furthermore, we denote the image of  $\Phi_D^{\mathcal{I}}$  in  $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$  by  $M_D^{\mathcal{I}}$ .

**Proposition 8** With the notation from Definition 2, the image  $M_D^{\mathcal{I}}$  is an embedded, smooth manifold in  $\mathcal{S} \times \mathcal{V} \times \mathcal{W}$ .

**Proof** We have to prove the following points:

- (i)  $\Phi_D^{\mathcal{I}}$  is smooth;
- (ii)  $\Phi_D^{\mathcal{I}}$  is injective (and hence a bijection onto its image);
- (iii) the inverse of  $\Phi_D^{\mathcal{I}}$  (on its image) is continuous;
- (iv)  $\Phi_D^{\mathcal{I}}$  is an immersion, that is, its derivative is injective everywhere.

Points (ii) and (iii) say that  $\Phi_D^{\mathcal{I}}$  is a *topological embedding*; points (i) and (iv) strengthen the result to show that  $\Phi_D^{\mathcal{I}}$  is even an embedding in the sense of differential geometry.

We will now give the (rather technical) proofs of the aforementioned four points. Throughout the proofs, we will always assume w.l.o.g. that the vertices of  $D = ([p], E)$  are numbered according to an inverse topological sorting, such that all matrices in  $\mathbf{B}(D)$  are strictly lower triangular matrices.

- (i) The smoothness of  $\Phi_D^{\mathcal{I}}$  is immediately clear from its definition:  $\Phi_D^{\mathcal{I}}$  is a composition of smooth functions.
- (ii) Let  $(B, \gamma)$  and  $(B', \gamma') \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$  such that  $\Phi_D^{\mathcal{I}}(B, \gamma) = \Phi_D^{\mathcal{I}}(B', \gamma')$ ; by the definition of  $\Phi_D^{\mathcal{I}}$ , this is the case if and only if  $K^{(I)}(B, \gamma) = K^{(I)}(B', \gamma')$  for all  $I \in \mathcal{I}$ . This condition simplifies to

$$(\mathbb{I} - B)R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{I} - B)^T = (\mathbb{I} - B')R^{(I)} \text{diag}(\gamma') R^{(I)} (\mathbb{I} - B')^T,$$

or, with the abbreviation  $A := (\mathbb{I} - B)^{-1}(\mathbb{I} - B')$ ,

$$R^{(I)} \text{diag}(\gamma) R^{(I)} A^{-T} = A R^{(I)} \text{diag}(\gamma') R^{(I)}. \quad (22)$$

By the assumption made before,  $B$  and  $B'$  are strict lower triangular matrices, hence  $A$  is a lower triangular matrix with ones as diagonal entries. Then, the left-hand side of equation (22) is an *upper* triangular matrix, whereas the right-hand side is a *lower* triangular matrix. We conclude that both sides of the equation must consist of a *diagonal* matrix, and that we can transpose the left-hand side:

$$A^{-1} R^{(I)} \text{diag}(\gamma) R^{(I)} = A R^{(I)} \text{diag}(\gamma') R^{(I)}. \quad (23)$$

For some  $a \notin I$ , the  $a^{\text{th}}$  column of the matrix equation (23) reads

$$(A^{-1} \text{diag}(\gamma))_{\cdot a} = (A^{-1})_{\cdot a} \gamma_a = A_{\cdot a} \gamma'_a = (A \text{diag}(\gamma'))_{\cdot a}. \quad (24)$$

Since the family of targets  $\mathcal{I}$  is *conservative*, there is, for every  $a \in [p]$ , some  $I \in \mathcal{I}$  such that  $a \notin I$ ; because equation 23 holds for every  $I \in \mathcal{I}$ , the column-wise equation (24) holds for every  $a \in [p]$ , so we finally find  $A^{-1} \text{diag}(\gamma) = A \text{diag}(\gamma')$ , or, equivalently,  $A^2 = \text{diag}(\gamma) \text{diag}(\gamma')^{-1}$ . Because the diagonal of  $A^2$  only consists of ones, we see that  $\gamma = \gamma'$ . It follows that  $A^2 = \mathbb{1}$ , and because  $A$  is a unit triangular matrix, this means that  $A = \mathbb{1}$ , and hence, by definition of  $A$ ,  $B = B'$ . Therefore,  $\Phi_D^{\mathcal{I}}$  is injective.

(iii) We can restrict our considerations to the parameterizations of the precision matrices:

$$K_{I^c, I^c}^{(I)} = P^{(I)} K^{(I)} P^{(I)\top} = (\mathbb{1} - B)_{I^c, I^c} \text{diag}(\gamma_{I^c}) (\mathbb{1} - B_{I^c, I^c})^\top, \quad (25)$$

$$\begin{aligned} K_{I, I^c}^{(I)} &= Q^{(I)} K^{(I)} P^{(I)\top} = -Q^{(I)} B R^{(I)} \text{diag}(\gamma) (P^{(I)\top} - R^{(I)} B^\top P^{(I)\top}) \\ &= -B_{I, I^c} \text{diag}(\gamma_{I^c}) (\mathbb{1} - B_{I^c, I^c})^\top. \end{aligned} \quad (26)$$

By assuming, as before, that  $B$  is a strict lower triangular matrix, (25) represents the Cholesky decomposition of  $K_{I^c, I^c}^{(I)}$ . This decomposition is unique, and  $B_{I^c, I^c}$  as well as  $\gamma_{I^c}$  depend *continuously* on  $K_{I^c, I^c}^{(I)}$  (Schwarz and Köckler, 2006).

For each  $b \in [p]$ , there is some  $I \in \mathcal{I}$  that does not contain  $b$  since  $\mathcal{I}$  is conservative. Hence  $\gamma_b$  can be calculated out of  $K_{I^c, I^c}^{(I)}$  by performing the Cholesky decomposition as described above. This shows that  $\gamma$  is a continuous function of the precision matrices  $(K^{(I)})_{I \in \mathcal{I}}$ .

Assume now that  $a \rightarrow b$  is an arrow in  $D$ , and let  $I \in \mathcal{I}$  be an intervention target with  $b \notin I$ . If  $a \notin I$ ,  $B_{ab}$  can also be calculated from  $K_{I^c, I^c}^{(I)}$  via the (continuous) Cholesky decomposition. Otherwise,  $B_{ab}$  is an entry of the matrix  $B_{I, I^c}$  which can be calculated by solving equation (26):

$$B_{I, I^c} = -K_{I, I^c}^{(I)} (\mathbb{1} - B_{I^c, I^c})^{-\top} \text{diag}(\gamma_{I^c})^{-1},$$

which is a *continuous* function since the matrix inversion is continuous. Altogether, also the matrix  $B \in \mathbf{B}(D)$  is a continuous function of the precision matrices  $(K^{(I)})_{I \in \mathcal{I}}$ , what proves the claim.

(iv) We have to show that the derivative  $d\Phi_D^{\mathcal{I}}(B, \gamma)$  has maximal rank for all  $(B, \gamma) \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$ . For that aim, we consider the canonical basis

$$\{(H^{(a,b)}, 0)\}_{(a,b) \in E} \cup \{(0, e_i)\}_{1 \leq i \leq p}$$

of  $\mathbf{B}(D) \times \mathbb{R}^p$ , the tangent space of  $\mathbf{B}(D) \times \mathbb{R}_{>0}^p$  at the point  $(B, \gamma)$ , where  $H^{(a,b)}$  denotes the  $p \times p$  matrix with  $H_{ab}^{(a,b)} = 1$  and  $H_{ij}^{(a,b)} = 0$  for  $(i, j) \neq (a, b)$ , and  $e_i$  denotes the  $i^{\text{th}}$  canonical basis vector of  $\mathbb{R}^p$ . We must show that

$$\{d\Phi_D^{\mathcal{I}}(B, \gamma)(H^{(a,b)}, 0)\}_{(a,b) \in E} \cup \{d\Phi_D^{\mathcal{I}}(B, \gamma)(0, e_i)\}_{1 \leq i \leq p}$$

is a linearly independent set for all  $(B, \gamma) \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$ . Again, it is sufficient to consider the derivatives of the precision matrices  $K^{(I)}$ .

We start with the directional derivative of  $K^{(I)}$  in direction  $(H^{(a,b)}, 0)$  for a pair  $(a, b) \in E$ . This derivative is

$$\begin{aligned} dK^{(I)}(B, \gamma)(H^{(a,b)}, 0) &= \\ &= H^{(a,b)} R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{1} - B)^\top - (\mathbb{1} - B) R^{(I)} \text{diag}(\gamma) R^{(I)} H^{(a,b)\top}. \end{aligned}$$

For a matrix  $A \in \mathbb{R}^{p \times p}$ , the matrix  $H^{(a,b)} A$  contains  $A_{b \cdot}$  as the  $a^{\text{th}}$  row; all other rows are filled with zeros. We then can see that

$$\left[ R^{(I)} \text{diag}(\gamma) R^{(I)} (\mathbb{1} - B)^\top \right]_{b \cdot} = \begin{cases} \gamma_b ((\mathbb{1} - B)_{\cdot b})^\top, & \text{if } b \notin I, \\ 0, & \text{otherwise.} \end{cases}$$

Based on these considerations and the fact that  $B$  is a strictly lower triangular matrix, one can then show that, if  $b \notin I$ ,  $\mathrm{d}K^{(I)}(B, \gamma)(H^{(a,b)}, 0) = F^{(a,b)}$ , where

$$F^{(a,b)} := \gamma_b \left( \begin{array}{c|ccc|ccc} & & & & 0 & & & \\ & & & & \vdots & & & \\ & & & & 0 & & & \\ & & & & -1 & & 0 & \\ & & & & B_{b+1,b} & & & \\ & & & & \vdots & & & \\ & & & & B_{a-1,b} & & & \\ \hline 0 & \cdots & 0 & -1 & B_{b+1,b} & \cdots & B_{a-1,b} & \\ \hline & & & & 2B_{ab} & & B_{a+1,b} & \cdots & B_{pb} \\ \hline & & & & B_{a+1,b} & & & & 0 \\ & & & & \vdots & & & & \\ & & & & B_{pb} & & & & \end{array} \right)$$

We continue with the calculation of the directional derivative of  $K^{(I)}$  in direction  $(0, e_b)$ ,  $1 \leq b \leq p$ . In this less tedious case, we that

$$\mathrm{d}K^{(I)}(B, \gamma)(0, e_b) = \begin{cases} (\mathbb{1} - B)_{\cdot b} ((\mathbb{1} - B)_{\cdot b})^T, & \text{if } b \notin I, \\ 0, & \text{otherwise.} \end{cases}$$

This means that, for  $b \notin I$ , we have  $\mathrm{d}K^{(I)}(B, \gamma)(0, e_b) = G^{(b)}$ , where

$$G^{(b)} := \left( \begin{array}{c|cccc} 0 & & & & 0 \\ \hline & 1 & -B_{b+1,b} & \cdots & -B_{pb} \\ 0 & -B_{b+1,b} & & & \\ & \vdots & & * & \\ & -B_{pb} & & & \end{array} \right)$$

It can easily be seen that the matrices  $\{F^{(a,b)}\}_{a>b} \cup \{G^{(b)}\}_{1 \leq b \leq p}$  are linearly independent. Since for each  $b \in [p]$ , there is some  $I \in \mathcal{I}$  with  $b \notin I$ , we can finally conclude that the set

$$\{\mathrm{d}\Phi_D^{\mathcal{I}}(B, \gamma)(H^{(a,b)}, 0)\}_{(a,b) \in E} \cup \{\mathrm{d}\Phi_D^{\mathcal{I}}(B, \gamma)(0, e_i)\}_{1 \leq i \leq p}$$

is linearly independent, which proves the claim.  $\square$

We have now shown that the parameter sets  $M_D^{\mathcal{I}}$  are smooth embedded manifolds. To be able to apply Theorem 6, it remains to show that two different parameter manifolds are not arbitrarily close.

**Proposition 9** *Let  $\mathcal{I}$  be a conservative family of targets, and let  $D_1$  and  $D_2$  be two DAGs that are not  $\mathcal{I}$ -equivalent. Assume that  $\theta \in \mathcal{S} \times \mathcal{V} \times \mathcal{W}$  is a parameter vector with  $\theta \in \overline{M_{D_1}^{\mathcal{I}}}$  and  $\theta \in M_{D_2}^{\mathcal{I}}$ . Then also  $\theta \in M_{D_1}^{\mathcal{I}}$  holds.*

**Proof** each  $j$ , a unique parameterization  $(B^{(j)}, \gamma^{(j)}) \in \mathbf{B}(D) \times \mathbb{R}_{\geq 0}^p$  such that  $\theta^{(j)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j)}, \gamma^{(j)})$ . The sequence  $(B^{(j)}, \gamma^{(j)})_{j \geq 1}$  must be bounded, otherwise the sequence  $\theta^{(j)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j)}, \gamma^{(j)})$  could not be bounded since  $K^{(I)}$ ,  $I \in \mathcal{I}$ , are polynomials in  $B$  and  $\gamma$  (Definition 2). By the theorem of Bolzano-Weierstrass we therefore have a subsequence  $(B^{(j_k)}, \gamma^{(j_k)})$  that converges to some  $(B, \gamma) \in \mathbf{B}(D) \times \mathbb{R}_{\geq 0}^p = \overline{\mathbf{B}(D) \times \mathbb{R}_{\geq 0}^p}$ .

The parameterization  $\Phi_{D_1}^{\mathcal{I}}$  has a continuous continuation on  $\mathbf{B}(D) \times \mathbb{R}_{\geq 0}^p$ . Therefore we have

$$\theta^{(j_k)} = \Phi_{D_1}^{\mathcal{I}}(B^{(j_k)}, \gamma^{(j_k)}) \xrightarrow{k \rightarrow \infty} \Phi_{D_1}^{\mathcal{I}}(B, \gamma) ,$$

and  $\Phi_{D_1}^{\mathcal{I}}(B, \gamma) = \theta$  holds because of the uniqueness of limits.

It remains to show that  $(B, \gamma) \in \mathbf{B}(D) \times \mathbb{R}_{>0}^p$ , that is, to show that  $\gamma_b \neq 0$  for all  $b \in [p]$ . Since  $\mathcal{I}$  is conservative, there is, for each  $b \in [p]$ , some  $I \in \mathcal{I}$  such that  $b \notin I$ . From Lemma 2, we know that

$$\det K^{(I)}(B, \gamma) = \det \tilde{K}^{(I)} \prod_{a \notin I} \gamma_a ;$$

since the prerequisite  $\theta \in M_{D_2}^{\mathcal{I}}$  implies  $\det K^{(I)} \neq 0$ , we conclude that  $\gamma_a \neq 0$  for all  $a \notin I$ . This in particular implies  $\gamma_b \neq 0$ , which completes the proof.  $\square$

We have now shown that the parameter sets  $M_D^{\mathcal{I}}$  of all DAGs  $D$  fulfill the prerequisites of Theorem 6; an immediate consequence is the following corollary:

**Corollary 10** *Consider model (6) with the family of intervention targets  $\mathcal{I}$ . Assume (A3) from Theorem 1, and the estimator*

$$\hat{D} = \arg \min_D -\ell_D(\hat{B}(D), \{\hat{\sigma}_k^2(D)\}_k; \mathcal{T}, \mathbf{X}) + \frac{1}{2} \log(n) \dim(D) .$$

*Then: as  $n \rightarrow \infty$ ,*

$$\mathbb{P}[\hat{D} \text{ is a minimum independence map}] \rightarrow 1 ,$$

*where  $\mathbb{P}$  refers to the probability distribution under the true model.*

As we noted in Section 3.2, every minimum independence map is  $\mathcal{I}$ -Markov equivalent to the true model if the true observational and all corresponding interventional densities are faithful. In this case (that is, under the assumptions (A1) and (A2) of Section 3.2), the optimization problem in (14) almost surely has a *unique* solution in the limit  $n \rightarrow \infty$ , namely the  $\mathcal{I}$ -Markov equivalence class of the true model (Theorem 1).

## References

- Andersson, S., D. Madigan, and M. Perlman (1997). A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics* 25, 505–541.
- Banerjee, O., L. El Ghaoui, and A. d’Aspremont (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research* 9, 485–516.
- Chickering, D. (2002a). Learning equivalence classes of Bayesian-network structures. *Journal of Machine Learning Research* 3, 445–498.
- Chickering, D. (2002b). Optimal structure identification with greedy search. *Journal of Machine Learning Research* 3, 507–554.
- Chickering, D. M. (1996). Learning Bayesian networks is NP-complete. In D. Fisher and H. Lenz (Eds.), *Learning from Data: Artificial Intelligence and Statistics V*, pp. 121–130. Springer.
- Cooper, G. F. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. In *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI 1999)*, pp. 116–125.



- Eaton, D. and K. Murphy (2007). Exact Bayesian structure learning from uncertain interventions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, Volume 2, pp. 107–114.
- Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics* 9, 432–441.
- Haughton, D. D. M. (1988). On the choice of a model to fit data from an exponential family. *The Annals of Statistics* 16(1), 342–355.
- Hauser, A. and P. Bühlmann (2012). Characterization and greedy learning of interventional Markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research* 13, 2409–2464.
- He, Y.-B. and Z. Geng (2008). Active learning of causal networks with intervention experiments and optimal designs. *Journal of Machine Learning Research* 9, 2523–2547.
- Hoyer, P., D. Janzing, J. Mooij, J. Peters, and B. Schölkopf (2009). Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems 21, 22nd Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pp. 689–696.
- Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research* 8, 613–636.
- Kalisch, M., M. Mächler, D. Colombo, M. Maathuis, and P. Bühlmann (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software* 47 (11), 1–26.
- Lauritzen, S. (1996). *Graphical Models*. Oxford University Press.
- Maathuis, M., M. Kalisch, and P. Bühlmann (2009). Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics* 37, 3133–3164.
- Meinshausen, N. and P. Bühlmann (2010). Stability selection (with discussion). *Journal of the Royal Statistical Society Series B* 72, 417–473.
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann.
- Pearl, J. (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press.
- Peters, J., J. Mooij, D. Janzing, and B. Schölkopf (2011). Identifiability of causal graphs using functional models. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*.
- Sachs, K., O. Perez, D. Pe’er, D. Lauffenburger, and G. Nolan (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science* 308(5721), 523–529.
- Schwarz, H. R. and N. Köckler (2006). *Numerische Mathematik* (6th ed.). Stuttgart: Vieweg + Teubner.
- Shimizu, S., P. Hoyer, A. Hyvärinen, and A. Kerminen (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research* 7, 2003–2030.
- Silander, T. and P. Myllymäki (2006). A simple approach for finding the globally optimal Bayesian network structure. In *Proceedings of the 22th Conference on Uncertainty in Artificial Intelligence (UAI 2006)*.

- Spirtes, P., C. Glymour, and R. Scheines (2000). *Causation, Prediction, and Search* (Second ed.). MIT Press.
- Tsamardinos, I., L. E. Brown, and C. F. Aliferis (2006). The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learning* 65(1), 31–78.
- Verma, T. and J. Pearl (1990). On the equivalence of causal models. In *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence (UAI 1990)*, pp. 220–227.